Gender Bias in Pretrained Swedish Embeddings

Magnus Sahlgren
RISE
Sweden
magnus.sahlgren@ri.se

Fredrik Olsson
RISE
Sweden
fredrik.olsson@ri.se

Abstract

This paper investigates the presence of gender bias in pretrained Swedish embeddings. We focus on a scenario where names are matched with occupations, and we demonstrate how a number of standard pretrained embeddings handle this task. Our experiments show some significant differences between the pretrained embeddings, with word-based methods showing the most bias and contextualized language models showing the least. We also demonstrate that a previously proposed debiasing method does not affect the performance of the various embeddings in this scenario.

1 Introduction

The motivation for this study is the currently widespread practice of using pretrained embeddings as building blocks for NLP-related tasks. More specifically, we are concerned about such usage by actors in the public sector, for instance government agencies and public organizations. It is obvious how the presence of (gender or racial) bias would be potentially serious in applications where embeddings are used as input to decision support systems in the public sector.

As an example, in Sweden limited companies must be approved and registered by the Swedish Companies Registration Office. One important (and internationally unique) step in this registration procedure is the approval of the company name, which is decided by case handlers at the Registration Office. Their decision is based on several factors, one of which is the appropriateness of the company name in relation to the company description. Now, imagine the hypothetical use case in which the case handlers use a decision support system that employs pretrained embeddings to quantify the similarity between a suggested company name and its company description. Table 1 exemplifies what the results might look like. In this fictive example, the company description states that the company will do business with cars, and the name suggestions are composed of a person name in genitive and the word "cars" (i.e. "Fredrik's cars"). We use pretrained Swedish ELMo embeddings (Che et al., 2018) to compute the distance between the name suggestion and the company description.

The results demonstrate that male person names ("Magnus" and "Fredrik") are closer to "cars" in the ELMo similarity space than female person names ("Maria" and "Anna"). If such results are used as input to a decision support system for deciding on the appropriateness of a company name suggestion in relation to a company description, we might introduce gender bias into the decision process. We subscribe to the view that such bias would be unfair and problematic.

The point of this paper is therefore to investigate gender bias when using existing and readily available pretrained embeddings for tasks relating to names and occupations. We include both word-based embeddings produced using

Name suggestion	Company description	Distance
Magnus bilar	Bolaget ska bedriva verksamhet med bilar	0.028
Fredriks bilar	Bolaget ska bedriva verksamhet med bilar	0.038
Marias bilar	Bolaget ska bedriva verksamhet med bilar	0.044
Annas bilar	Bolaget ska bedriva verksamhet med bilar	0.075

Table 1: Examples of gender bias with respect to occupations using pretrained ELMo embeddings.

word2vec and fastText, as well as character-based (and WordPiece-based) contextualized embeddings produced using ELMo and the multilingual BERT. The next section covers related work. We then discuss the various embeddings in Section 3, before we then turn to some experimental evidence of bias in the embeddings, and we also show that the previously proposed debiasing method is unable to handle gender bias in our scenario.

2 Related work

Research regarding bias and stereotypes expressed in text and subsequently incorporated in learned language models is currently a vivid field. Caliskan et al. (2017) show that learned embeddings exhibit every linguistic bias documented in the field of psychology (such as that flowers are more pleasant than insects, musical instruments are preferred to weapons, and personal names are used to infer race). Garg et al. (2018) show that temporal changes of the embeddings can be used to quantify gender and ethnic stereotypes over time, and Zhao et al. (2017) suggest that biases might in fact be amplified by embedding models.

Several researchers have also investigated ways to counter stereotypes and biases in learned language models. While the seminal work by Bolukbasi et al. (2016a, 2016b) concerns the identification and mitigation of gender bias in *pretrained* word embeddings, Zhao et al. (2018) provide insights into the possibilities of *learning* embeddings that are gender neutral. Bordia and Bowman (2019) outline a way of training a recurrent neural network for word-based language modelling such that the model is gender neutral. Park et al. (2018) discuss different ways of mitigating gender bias, in the context of abusive language detection, ranging from debiasing a model by using the hard debiased word embeddings produced by Bolukbasi et al. (2016b), to manipulating the data prior to training a model by swapping masculine and feminine mentions, and employing transfer learning from a model learned from less biased text.

Gonen and Goldberg (2019) contest the approaches to debiasing word embeddings presented by Bolukbasi et al. (2016b) and Zhao et al. (2018), arguing that while the bias is reduced when measured according to its definition, i.e., dampening the impact of the general gender direction in the vector space, "the actual effect is mostly hiding the bias, not removing it". Further, Gonen and Goldberg (2019) claim that a lot of the supposedly removed bias can be recovered due to the geometry of the vector representation of the gender neutralized words.

Our contribution consists of an investigation of the presence of gender bias in pretrained embeddings for Swedish. We are less interested in bias as a theoretical construct, and more interested in the effects of gender bias in actual applications where pretrained embeddings are employed. Our experiments are therefore tightly tied to a real-world use case where gender bias would have potentially serious ramifications. We also provide further evidence of the inability of the debiasing method proposed by Bolukbasi et al. (2016b) to handle the type of bias we are concerned with.

3 Embeddings

We include four different standard embeddings in these experiments: word2vec, fastText, ELMo and BERT. There are several pre-trained models available in various web repositories. We select one representative instance per model, summarized in Table 2 (next page).

These models represent different types of embeddings. word2vec (Mikolov et al., 2013) builds embeddings by training a shallow neural network to predict a set of context words based on a target word (this is the so-called *skipgram* architecture; if we instead predict the target word based on the context words the model is called *continuous bag of words*). The network learns two sets of vectors, one for the target terms (the embedding vectors), and one for context terms. The objective of the network is to learn vectors such that their dot product correspond to the log likelihood of observing word pairs in the training data. fastText (Bojanowski et al., 2017) uses the same neural network architecture, but incorporates character information by using character n-grams instead of whole words in the prediction step.

It should be noted that most applications of the above-mentioned vectors use only the embeddings for the target terms. In fact, many repositories with pretrained vectors do not even contain the context embeddings. When the downstream task focuses on *associative* relations (which is the case in the present scenario with names and occupations), it would be beneficial to be able to use *both* target and context vectors, since using only one of these will result in more *paradigmatic* similarities.

Model	Source Code Repository	Training data
word2vec	vectors.nlpl.eu	CoNLL17 data
fastText	github.com/facebookresearch/fastText	Wikipedia
ELMo	github.com/HIT-SCIR/ELMoForManyLangs	CoNLL18 data
BERT	github.com/google-research/bert	Wikipedia

Table 2: The pre-trained embeddings and models included in these experiments were downloaded in April 2019 from the following URLs. **word2vec**: vectors.nlpl.eu/repository/11/69.zip, **fastText**: dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.sv.300.bin.gz, **ELMo**: vectors.nlpl.eu/repository/11/173.zip, **BERT**: storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip

ELMo (Peters et al., 2018) is a deep character-based neural network that learns embeddings by predicting the next token given an input sequence. The network architecture includes both convolutional and (bidirectional) LSTM layers, and produces an embedding that that is sensitive to the particular context of the input sequence. ELMo is thus different from word2vec and fastText in the sense that it produces *contextualized* embeddings, which has proven to be highly beneficial when using the embeddings as representation in downstream NLP tasks such as classification, entity recognition, and question answering.

BERT (Devlin et al., 2018) is similar to ELMo in the sense that it uses a deep neural network architecture and produces contextualized embeddings. However, it differs in the type of network used. BERT uses a (bidirectional) Transformer network that relies exclusively on attention, and the model is trained using a masked language model task, similar to a cloze test. Contrary to ELMo, BERT is not character-based, but relies on WordPiece tokenization of the input data. This has some potentially problematic effects when tokenizing proper names. As an example, the Swedish male name "Henrik" gets tokenized as ["hen", "##rik"], with "rik" probably deriving from the Swedish word "rik" (eng. "rich"). It would have been desirable to *not* use WordPiece tokenization for proper names.

In the following experiments, pre-trained ELMo and BERT are used to produce contextualized embeddings both for individual words (such as names or places) and for texts (such as company descriptions). Pre-trained word2vec and fastText are used to look up individual words, and for texts we follow standard practice and average the vectors of the component words. Since proper names in Swedish use uppercase for the initial letter, we retain the casing information for all models that can handle such vocabulary, which in our case are all models except word2vec.

4 Data

In order to investigate whether our concerns about gender bias in pretrained Swedish embeddings are valid, we collect lists of the 100 most common Swedish female and male first names from Statistics Sweden (www.scb.se). We also collect lists of the most typical female and male occupations from the same source, as shown in Tables 3 and 4 (next page). These are the most common occupations for women and men as compiled by Statistics Sweden, together with the percentage of women and men in each occupation.

Since our interest in this paper is bias, we do not include occupations that have less than (or close to) 50% occurrence of women or men (such cases are marked by $*$ in the tables). This leaves us with 18 typically female occupations, and 15 typically male occupations. Some of the remaining occupations are very similar to each other, and we therefore collapse them to one occupation (marked by numbers in the tables), resulting in 14 distinct female occupations and 14 distinct male occupations. For each of these *gendered occupations*, we also list a number of synonyms, collected from wikipedia.se and framtid.se. Morphological variants of each term are included.

5 Experiment 1: names and occupations

As a first experiment, we compute the similarity between the names and the occupations using the different embeddings. We do this by computing the similarity between each name and each occupation. Table 5 shows the percentage of female and male names that are on average more similar to a female vs. male occupation. Numbers in parentheses are based on only *the most similar* oc-

Occupation (Swedish)	Occupation (English)	% women
[1]Undersköterska	Assistant nurse	92
Barnskötare	Nanny	89
Grundskollärare	Primary school teacher	75
Förskollärare	Preschool teacher	96
[2]Butikssäljare, fackhandel	Shop sales	61
[3]Vårdbiträde	Care assistant	81
Kontorsassistent och sekreterare	Secretary	79
Städare	Cleaner	75
Personlig assistent	Personal assistant	74
[2]Butikssäljare, dagligvaror	Retail sales	67
[3]Vårdare, boendestödjare	Housing assistant	73
Restaurang- och köksbiträde	Restaurant assistant	65
Planerare och utredare	Planner	63
Grundutbildad sjuksköterska	Nurse	90
[4]Ekonomiassistent	Accountant assistant	88
[1]Undersköterska, vård- och specialavdelning	Nursing staff	91
∗ Företagssäljare	Company sales	27
∗ Kock och kallskänka	Chef	52
[4]Redovisningsekonomer	Accountant	79
Socialsekreterare	Social worker	86

Table 3: The 20 most common occupations for Swedish women in 2016 according to Statistics Sweden (`www.scb.se`).

Occupation (Swedish)	Occupation (English)	% men
Företagssäljare	Company sales	73
Lager- och terminalpersonal	Warehouse staff	79
Mjukvaru- och systemutvecklare	Software developer	80
Lastbilsförare	Truck driver	94
Träarbetare, snickare	Carpenter	99
Maskinställare och maskinoperatörer	Machine operator	86
∗ Butikssäljare, fackhandel	Shop sales	39
Fastighetsskötare	Janitor	86
Motorfordonsmekaniker och fordonsreparatör	Vehicle mechanic	97
Installations- och serviceelektriker	Electrician	98
∗ Butikssäljare, dagligvaror	Retail sales	33
∗ Grundskollärare	Primary school teacher	25
Underhållsmekaniker och maskinreparatör	Maintenance mechanic	95
∗ Planerare och utredare	Planner	37
∗ Restaurang- och köksbiträde	Restaurant assistant	35
[1]Ingenjör och tekniker inom elektroteknik	Electrical technician	87
[1]Civilingenjörsyrke inom elektroteknik	Electrical engineer	84
Verkställande direktör	CEO	84
Buss- och spårvagnsförare	Bus driver	86
VVS-montör	Plumber	99

Table 4: The 20 most common occupations for Swedish men in 2016 according to Statistics Sweden (`www.scb.se`).

Model	Male names Male occupations	Male names Female occupations	Female names Male occupations	Female names Female occupations
word2vec	91 (86)	9 (14)	99 (98)	1 (2)
fastText	4 (10)	96 (90)	100 (100)	0 (0)
ELMo	96 (63)	4 (37)	49 (87)	51 (13)
BERT	37 (54)	63 (46)	76 (55)	24 (45)

Table 5: Percentage of female and male names that are on average more similar to a female vs. male occupation. The similarities are calculated based on the original embeddings, before the application of the debiasing step described in Section 6. Numbers in parentheses only count the single most similar occupation for each name.

cupation for each name. As an example, imagine we only have two female and male occupations, and that the name "Anna" has the similarities 0.47 and 0.78 to the female occupations, and the similarities 0.12 and 0.79 to the male occupations. In this example, "Anna" would be closer to the female occupations when counting the average similarities (0.625 vs. 0.455), but closer to the male occupations when only considering the most similar examples (0.79 vs. 0.78).

There are several ways in which an embedding could show bias in this setting. The arguably most detrimental effect would be if the embedding grouped male names with male occupations and female names with female occupations. Somewhat less severe, but still problematic, would be if the embedding grouped all names with female or male occupations. A completely unbiased model would not show any difference between the female and male names with respect to female and male occupations.

The numbers in Table 5 demonstrate some interesting differences between the different embeddings. word2vec shows a clear tendency to group both male and female names with male occupations. fastText, on the other hand, shows a bias for female occupations for male names, and for male occupations for female names. This is a very interesting difference, given that the only algorithmic difference between these models is the inclusion of character n-grams in the latter model.

The results for ELMo and BERT show some interesting differences too. ELMo groups the male names with the male occupations, but is less biased for the female names. When counting only the single most similar occupation, ELMo shows a similar tendency as word2vec and groups both male and female names with male occupations. BERT, on the other hand, seems slightly more balanced, with a tendency similar to fastText when counting the average similarities. When only counting the single most similar occupation, BERT is almost perfectly balanced between female and male occupations.

6 Debiasing embeddings

We apply the debiasing methodology in (Bolukbasi et al., 2016b) to the pretrained embedddings. Debiasing a given vector space involves finding the general direction in it that signifies gender using a set of predefined *definitional pairs*, and then removing the direction from all vectors except those corresponding to words that are naturally *gender specific*.

The definitional pairs are word pairs expressing among themselves a natural distinction between the genders, e.g., *he – she*, and *mother – father*. In our setting, there are 10 such pairs. The gender specific words are words that also carry a natural gender dimension that should not be corrected during the debiasing phase of the vector space. We use the same methodology for growing a seed set of gender specific words into a larger set as described in (Bolukbasi et al., 2016b), and end up with 486 manually curated gender specific words, including e.g., *farfar* (paternal grandfather), *tvillingsystrar* (twin sisters), and *matriark* (matriarch).

The definitional pairs are used to find a *gender direction* in the embedding space, which is done by taking the difference vector of each of the definitional pairs (i.e. $w_1 - w_2$), and then factorizing the mean-centered difference vectors using PCA, retaining only the first principal component, which will act as the gender direction. The vector space is then *hard debiased*[1] in the sense that the gen-

[1]The alternative is *soft debiasing*, in which one tries to strike a balance between keeping the pairwise distances

Model	Male names Male occupations	Male names Female occupations	Female names Male occupations	Female names Female occupations
word2vec	88 (89)	12 (11)	95 (93)	5 (7)
fastText	0 (10)	100 (90)	100 (99)	0 (1)
ELMo	99 (87)	1 (13)	26 (71)	74 (29)
BERT	0 (50)	100 (50)	97 (52)	3 (48)

Table 6: Percentage of female and male names that are on average more similar to a female vs. male occupation. The similarities are calculated based on the debiased version of each model. Numbers in parentheses only count the single most similar occupation for each name.

der direction b is removed from the embeddings of all non-gender specific words w using orthogonal projection: $w' = w - b \times \frac{w \cdot b}{b \cdot b}$.

The approach described by (Bolukbasi et al., 2016b) includes an *equalize* step to make all gender neutral words equidistant to each of the members of a given *equality* set of word pairs. The equality set is application specific, and since the current investigation of Swedish language embeddings does not naturally lend itself to include an equality set, the debiasing of the embeddings does not involve equalization in our case.

We apply the method described above to all pretrained embeddings in Table 3, as well as to the token vectors generated by ELMo and BERT. Although it is not clear whether the proposed debiasing method is applicable to embeddings produced by contextualized language models, we argue that it is reasonable to treat the contextualized models as black boxes, and rely only on their output, given the proposed use case.

7 Experiment 2: names and occupations (revisited)

We repeat the experiment described in Section 5, but using the debiased embeddings. Table 6 summarizes the results. It is clear that the debiasing method does not have any impact on the results in these experiments. The tendencies for the word-based embeddings word2vec and fastText are more or less identical before and after debiasing. The most striking differences between Table 5 and Table 6 are the results for ELMo and BERT, which become less balanced after applying the debiasing method. ELMo actually shows a clearer gender distinction *after* debiasing, with male names being more similar to male occupations, and female names being more similar to female occupations. BERT also becomes less balanced *after* debiasing, grouping male names with female occupations, and female names with male occupations, when considering the average similarities. When counting only the most similar occupation per name, BERT is still well balanced after debiasing.

8 Experiment 3: company names and company descriptions

The experiments in the previous sections are admittedly somewhat simplistic considering the scenario discussed in the Introduction: quantifying the similarity between a company name and a company description. In particular the contextualized language models are not primarily designed for generating token embeddings, and it is neither clear what kind of quality we can expect from such un-contextualized token embeddings, nor whether they are susceptible to the debiasing operation discussed in Section 6. In order to provide a more realistic scenario, we also include experiments where we compute the similarity between a set of actual company descriptions and a set of fictive company names generated from the lists of male and female names by adding the term "Aktiebolag" (in English limited company) after each name.[2]

The company descriptions are provided by the Swedish Companies Registration Office, and contain approximately 10 company descriptions for each of the sectors *construction work*, *vehicles and transportation*, *information technologies*, *health and health care*, *education*, and *economy*. Based on Tables 3 and 4, we consider the descriptions from the first three sectors to be representative of typically male occupations, and the descriptions from the latter three sectors to be representative

among all vectors and decreasing the influence of the gender specific direction.

[2]It is not uncommon for names of limited companies (in Sweden) to feature a person name and the term "Aktiebolag".

Model	Male names Male occupations	Male names Female occupations	Female names Male occupations	Female names Female occupations
word2vec	29 (29)	71 (71)	30 (30)	70 (70)
fastText	60 (61)	40 (39)	60 (61)	40 (39)
ELMo	52 (53)	48 (47)	53 (54)	47 (46)
BERT	42 (40)	58 (60)	41 (41)	59 (59)

Table 7: Percentage of female and male names that are on average more similar to a female vs. male occupation. The similarities are calculated based on the original embeddings, using the names and occupations in context. Numbers in parentheses only count the single most similar occupation for each name.

of typically female occupations.

We generate vectors for each of the descriptions and for each fictive company name (i.e. a male or female name, followed by "Aktiebolag"). For the word-based models (word2vec and fastText), we take the average of the embeddings of the words in the descriptions and the name. For the contextualized language models (ELMo and BERT), we generate vectors for each description and each fictive name. In the case of ELMo we take the average over the three LSTM layers, and for BERT we use the output embedding for the [CLS] token for each of the input sequences.

The results are summarized in Table 7. It is clear that these results are significantly more balanced than the results using tokens only. Even so, there are still some interesting differences between the embeddings. Contrary to the results in Tables 5 and 6, word2vec now shows a bias for female occupations, and fastText now shows a bias for male occupations. ELMo and BERT seem more balanced, with ELMo showing almost perfectly balanced results, and BERT showing a slight bias for female occupations.

Even though the biases apparently are different when considering tokens in comparison with considering texts, there *are* still biases in all models in both cases. The only exception in our experiments is ELMo, when used for texts instead of tokens. We hypothesize that the results for BERT are negatively affected by artefacts of the WordPiece tokenization, as discussed in Section 3.

9 The effect of debiasing on embeddings

So far, we have shown that all Swedish pretrained embeddings included in this study exhibit some degree of gender bias when applied to a real-world scenario. We now turn to investigate the effect the hard debiasing operation has on the embedding spaces, using the intrinsic evaluation methodology of Bolukbasi et al. (2016b). In this setting, a number of analogy pairs are extracted for the original and debiased embeddings, and human evaluators are used to asses the number of appropriate and stereotypical pairs in the respective representations. Bolukbasi et al. (2016b) used 10 crowdworkers to classify the analogy pairs as being appropriate or stereotypical. Their results indicated that 19% of the top 150 analogies generated using the original embedding model were deemed gender stereotypical, while the corresponding figure for the hard debiased model was 6%.

We carry out a similar, but smaller, evaluation exercise using the analogy pairs generated by the original Swedish word2vec and fastText models, as well as their debiased counterparts.[3] We use *hon – han* (*she – he*) as seed pair, and score all word pairs in the embeddings with respect to the similarity of the word pair's difference vector to that of the the seed pair. The top 150 pairs are manually categorized as either *appropriate*, *gender stereotypical*, or *uncertain* by the authors.

The results of the annotation are shown in Table 8 (next page). Due to the limited extent of the evaluation, we can only use these results for painting the big picture. First of all, there is a relatively small overlap between the analogy pairs in the top 150 of the original models, and the top lists of the debiased models: for word2vec, only 42 of the analogy pairs in the original list are also in the list produced by the debiased model. The corresponding number for fastText is 31. This means that the debiasing operation changes

[3] It would have been preferable to also include ELMo and BERT in this experiment, but generating vectors for large vocabularies using these models takes a prohibitively long time, and it is neither clear whether the resulting token embeddings make sense, not whether the debiasing operation is applicable to the resulting embeddings.

Analogies quality	Original word2vec	Debiased word2vec	Original fastText	Debiased fastText
Appropriate	97	52	135	36
Stereotypical	3	13	5	4
Uncertain	18	36	0	45

Table 8: The number of appropriate, stereotypical, and uncertain analogies in the top 150 pairs for the original and debiased embeddings. The numbers are the analogy pairs for which the annotators agree on the category.

the embedding space to a large extent. Second, there is a considerable amount of annotator uncertainty involved, either regarding the plausibility of a given analogy pair, or regarding its appropriateness. This is manifested by an increase of the number of uncertain analogy pairs that the annotators agree on between the original and debiased models (both for `word2vec` and `fastText`). However, the most interesting findings have to do with the number of stereotypical analogy pairs. The number of stereotypical analogy pairs output by the Swedish models is small compared to the numbers reported by Bolukbasi et al. (2016b). Further, the number of stereotypical pairs is *larger* in the debiased `word2vec` model than in the original model (we anticipated that it should be lower). It thus seems as if the debiasing operation makes the `word2vec` embedding space *more* biased. For `fastText`, the number of such pairs are slightly fewer in the debiased model compared to its original counterpart.

10 Discussion

This paper has shown that pretrained Swedish embeddings *do* exhibit gender bias to varying extent, and that the debiasing operation suggested by Bolukbasi et al. (2016a) does not have the desired effect, neither in the task of matching person names with occupations, nor in the case of the gender stereotypes being present among the top ranked analogy pairs generated by the models. Our experiments also indicate that word-based embeddings are more susceptible to bias than contextualized language models, and that there is an unexpectedly large difference in the biases shown by `word2vec` and `fastText`, something we believe requires further study.

Although contextualized language models appear to be more balanced with respect to gender bias in our experiments, there *is* still bias in these models; in particular if they are used to generate token embeddings, but also when they are used to generate representations for texts – `ELMo`, which produces almost perfect scores in Table 7, may still show bias in individual examples, such as those in Table 1. We acknowledge the possibility that it may not be appropriate to use contextualized language models to generate embeddings for individual tokens, but we also believe such usages to occur in real-world applications, and we therefore consider it relevant to include such examples in these experiments.

The debiasing operation proposed by Bolukbasi et al. (2016a) does nothing to rectify the situation in our setting. On the contrary, the debiased models still show significant gender bias, and in the case of `ELMo` and `BERT`, the bias actually becomes more prevalent *after* debiasing. (However, we are aware that the debiasing operation may neither be intended nor suitable for such representations.) Furthermore, our (admittedly small) analogy evaluation shows that debiasing actually introduces *more* stereotypical word pairs in the `word2vec` model.

Why then does not debiasing the Swedish word-based embeddings produce results similar to those of Bolukbasi et al. (2016a)? One of the big differences between the Swedish pretrained `word2vec` model and the one used by Bolukbasi et al. is the size of the vocabulary. The Swedish model contains 3M+ word types, while Bolukbasi et al. constrained their experiments to include only lowercased words shorter than 20 characters, omitting digits and words containing punctuation, from the top 50,000 most frequent words in the model. By doing so, Bolukbasi et al. effectively removed many person names from the model. A large portion of the word pairs in our analogy lists produced by the original model consist of person names (e.g., *Anna – Jakob*), which we consider to be *appropriate*, and their presence on the top 150 list contribute to the comparatively low number of

stereotypical pairs. The debiasing operation of the word-based models remove many of the persons name pairs from the top list, giving way for potentially stereotypical pairs. Thus, the increase of stereotypical pairs on the top list of analogy pairs generated by a debiased model is more likely to be due to the debiasing operation effectively removing many of the names from the top list, than the model being more biased in the first place.

Since our experiments have focused on pretrained embeddings readily available on the Internet, which have been trained on different types and different sizes of data, we cannot speculate about the extent to which a particular learning algorithm amplifies or distorts bias. We believe this is an interesting direction for further research, and we aim to replicate this study using a variety of embeddings trained on the same data.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016a. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016b. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *NAACL Student Research Workshop*.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.